REMARKS

Reconsideration is requested.

Claims 69-111 are pending.

The specification has been amended to include the attached formal drawings. Approval of the same is requested in the Examiner's next Action.

The title has been amended as suggested by the Examiner.

The specification has been amended to include further copies of pages 18-24, as requested by the Examiner in paragraph 8 of the Office Action dated November 19, 2002 (Paper No. 11). Moreover, the disclosure has been amended at page 349, line 25, as requested by the Examiner.

Withdrawal of the objections to the specification noted in paragraphs 7-9 of Paper No. 11 is requested.

The Section 112, first paragraph, rejection of claims 79-83 is traversed. Reconsideration and withdrawal of the rejection are requested in view of the following comments.

According to the molecular genetic classification method, when two microorganisms having DNAs which are strongly hybridizable with each other in a DNA-DNA hybridization experiment using chromosomal DNAs, they can be classified into the same genera. Accordingly, it will be apparent to one of ordinary skill in the art that microorganisms belonging to the genus *Corynebacterium* have a high percentage homology in chromosomal DNA.

Furthermore, *Corynebacterium glutamicum* is a species allied to microorganisms belonging to the genus *Brevibacterium* or the genus *Microbacterium*, as shown in the homology of the nucleotide sequences of rDNA (see the Attachment 2-4 page "Comparison in similarity of 16S-RNA of Corynebacterium glutamicum with microorganisms belonging to the genus Microbacterium or Brevibacterium"). Additionally, since there is a high degree of identity and similarity of the nucleotide sequences among the genes in the microorganisms, it is apparent that the chromosomal DNAs have a high homology.

Accordingly, the invention recited in claims 79-83 are submitted to be supported by an enabling disclosure.

The applicants believe that even if the identity in the amino acid sequences as a whole of two proteins is 30%, there is a high probability that both proteins have the same function so long as the similarity is about 60%. Furthermore, even at around 30% of both of the identity and similarity, so long as it is known that genes in a data base which are considered to be similar to target genes is known to be one of gene groups forming a cluster on the chromosomal DNA, whether the function of the protein encoded by the target genes has the same function as the protein encoded by the genes in the data base can be judged to a great degree by comparing the identify and similarity of proteins encoded by genes present in the upstream and downstream of the respective gene. Moreover, even at a low similarity in proteins as a whole, so long as a partial sequence which is assumed to be a domain of an active site or various functions has a high identity and similarity, the function of the protein encoded by the target genes can be identified at a high accuracy. For example, in BLAST which is a software program

for homology search used for presuming the function of a protein exemplified in the present application, a similar protein is not extracted based on the identity as a whole, but a sequence having a low e-value is extracted as a sequence having a high homology from partial sequences having a high homology score. Accordingly, even if the homology as a whole is around 30%, the function presumption is not led to a low accuracy the function presumption (see the Attachments 1 (12 page NCBI publication "The Statistics of Sequence Similarity Scores") and 2).

As discussed above, one of the ordinary skill would easily understand that the functions of all genes and proteins in *Corynebacterium glutamicum* carried out in the present invention are not determined at a low accuracy.

Furthermore, *Corynebacterium glutamicum* is a species allied to microorganisms belonging to the genus *Brevibacterium* or the genus *Microbacterium*, as shown in the homology of the nucleotide sequences of rDNA (see the Attachment 2). Additionally, since the identity and similarity of the nucleotide sequences among the genes in the microorganisms, it is apparent that the chromosomal DNAs have a high homology.

Accordingly, when the function of a protein encoded by a gene of a microorganism belonging to the genus *Brevibacterium* or the genus *Microbacterium* is examined, it would be apparent for one of ordinary skill in the art that the function can be presumed at a high accuracy by using a data base having nucleotide sequences and amino acid sequences of *Corynebacterium glutamicum* having a higher identity and homology with the respective gene and protein, without presumption from the function of genes registered in GenBank, *etc.* which are quite different in genus.

Therefore, one of ordinary skill in the art would analyze the function of a target gene and protein of a coryneform bacterium without undue experimentation based on the description in the present specification.

The Examiner asserts that sequences are homologous if they are related by divergence from a common ancestor, and that conversely, analogy relates to the acquisition of common structural of functional features via convergent evolution from unrelated ancestors.

Furthermore, the Examiner asserts that there are proteins which share no sequence or functional similarity, despite their common architecture. Moreover, the Examiner asserts that there are proteins which share groups of catalytic residues with almost identical spatial geometries, but they have no other sequence or structural similarities. Additionally, the Examiner asserts that the unreliability of predicting function based on similar sequences could be attributed to numerous annotation errors. See, paragraph 4 of Paper No. 11.

In the present invention, all nucleotide sequences of the chromosomal DNA of *Corynebacterium glutamicum* were discovered and all ORF's were identified. In the sequence data bases present at the time the present application was filed, nucleotide sequences, *etc.* having a gene group of various genes, e.g., gene group having a cluster structure, with a length to some extent, were registered, and there were many sequences of which positions in the chromosomal DNA were known.

When the functions of ORFs appearing in the present application are determined, the identity and similarity of a protein encoded by each ORF should be analyzed, and the function of the protein can be presumed at a high accuracy based on the

relationship with the function presumption of proteins encoded by the upstream and downstream of the ORF and its surrounding ORFs, in addition to the identity and similarity analyses.

For example, when a gene group relating to the metabolism of a compound forms a cluster on the chromosomal DNA is known, the probability that genes unrelating to the metabolism of the compound are present in the cluster would be considered to be low. Accordingly, when structure motif search is carried out for a protein encoded by a gene which is expected to be present in the cluster, a function which is not related to the metabolism is not determined for the gene even if a protein which has a similar structural motif but has a different function which is not related to the metabolism of the compound is extracted in the structural motif search. Thus, as a result of all nucleotide sequences of the chromosomal DNA being determined and all ORFs being identified, the accurate function can be determined without undue experimentation.

Based on the above, the determined functions of the proteins in the present invention are sufficiently reliable such that one of ordinary skill in the art would expect to determine according to the claimed invention the function of a target protein of a coryneform microorganisms having proteins which clearly have high homology and similarity to proteins of *Corynebacterium glutamicum* without undue experimentation based on the description in the present specification.

Withdrawal of the Section 112, first paragraph, rejection of claims 79-83 is requested.

The Section 112, second paragraph, rejection of claims 79-82 is, to the extent not obviated by the above, traversed. Reconsideration and withdrawal of the rejection are requested in view of the following comments.

The phrase "coincident with or analogous to" between the target amino acid sequence and the amino acid sequence in a data base will be recognized by one of ordinary skill in the art to mean that, for example, an e-value is e^{-10} or less, for example, when FASTA is used as a homology search program (page 349, lines 1-12 in the present specification). When a program other than FASTA is used, one of ordinary skill in the art would easily understand the index value in the program corresponding to "e-value = e^{-10} or less" when FASTA is used, based on the manufacture's instructions of the program.

The structure motif can be identified by the amino acid sequence in which a specific amino acid appears at the definite interval. Thus, the value based on the "coincident with or analogous to" can be fixed in the same manner as the case where the amino acid sequence is used as a target sequence.

Accordingly, the phrase "coincident with or analogous to" would be considered to be definite by one of ordinary skill in the art.

The Examiner's concerns relating to the definiteness of the recited comparator are unfounded. The programs used in the presently claimed methods for comparisons are well known and widely available. Programs used for methods of homology searching in the presently claimed invention include programs described, for example, at page 74, line 19 to page 75, line 3 in the present specification. These programs often contain conversion programs which also allow conversion from a nucleotide

sequence to an amino acid sequence and from an amino acid sequence to a nucleotide sequence. Therefore, one of ordinary skill in the art would recognize that selection of a particular homology search program may be required and carry out the desired homology analyses without requiring further definition. The metes and bounds of the claimed invention are clear in this regard.

As for the Examiner's concerns expressed in paragraph 9 of Paper No. 11, the applicants believe the claims, as a whole, define the invention and infringement is measured by the claimed invention. The Examiner's dissection of the claims between "preamble" and "active steps" is not believed, with all due respect, to be relevant to the issue of whether the claims are definite.

The claims are submitted to be definite and withdrawal of the Section 112, second paragraph, rejection is requested.

The Section 102 rejection of claims 79-83 over Mewes (Nucleic Acid Research (1998) Vol. 26, 1) is traversed. Reconsideration and withdrawal of the rejection are requested as, by the Examiner's own admission, the cited document fails to teach each and every aspect of the presently claimed invention. Specifically, the Examiner admits that page 8 of Paper No. 11 that the reference does not teach the recited elements of the claim in that the reference fails to teach or suggest the recited sequences. With regard to the Examiner's reliance on MPEP §2106 (IV)(B)(b), the applicants note that the presently claimed invention provides more than mere data or sequences located in a data storage means as non-functional descriptive material. Specifically, claim 79 provides a system which includes a user input device, data storage device, a computer based comparator which acts on the data through at least one program and an output

device. Claim 80 provides a method with an active recitation of method steps. Claim 81 provides a system which also requires a user input device and other aspects similarly provided by claim 79. Claim 81 provides similar required elements. Claim 82 provides a method which requires an active recitation of method steps. Claim 83 includes a data storage device or recording medium containing information as well as an executable computer program. Accordingly, the presently claimed invention defines statutory subject matter which is not taught or suggested in the cited art. Withdrawal of the Section 102 rejection is requested.

The claims are submitted to be in condition for allowance a Notice to that effect is requested.

Respectfully submitted,

NIXON & VANDERHYE P.C.

By: _____

B. J. Sadoff
Reg. No. **36,663**

1100 North Glebe Road, 8th Floor
Arlington, VA 22201-4714
Telephone: (703) 816-4000
Facsimile: (703) 816-4100

738036

The Statistics of Sequence Similarity Scores

Introduction

To assess whether a given alignment constitutes evidence for homology, it helps to know how strong an alignment can be expected from chance alone. In this context, "chance" can mean the comparison of (i) real but non-homologous sequences; (ii) real sequences that are shuffled to preserve compositional properties [1-3]; or (iii) sequences that are generated randomly based upon a DNA or protein sequence model. Analytic statistical results invariably use the last of these definitions of chance, while empirical results based on simulation and curve-fitting may use any of the definitions.

The statistics of global sequence comparison

Unfortunately, under even the simplest random models and scoring systems, very little is known about the random distribution of optimal global alignment scores [4]. Monte Carlo experiments can provide rough distributional results for some specific scoring systems and sequence compositions [5], but these can not be generalized easily. Therefore, one of the few methods available for assessing the statistical significance of a particular global alignment is to generate many random sequence pairs of the appropriate length and composition, and calculate the optimal alignment score for each [1,3]. While it is then possible to express the score of interest in terms of standard deviations from the mean, it is a mistake to assume that the relevant distribution is normal and convert this *Z*-value into a *P*-value; the tail behavior of global alignment scores is unknown. The most one can say reliably is that if 100 random alignments have score inferior to the alignment of interest, the *P*-value in question is likely less than 0.01. One further pitfall to avoid is exaggerating the significance of a result found among multiple tests. When many alignments have been generated, e.g. in a database search, the significance of the best must be discounted accordingly. An alignment with *P*-value 0.0001 in the context of a single trial may be assigned a *P*-value of only 0.1 if it was selected as the best among 1000 independent trials.

The statistics of local sequence comparison

Fortunately statistics for the scores of local alignments, unlike those of global alignments, are well understood. This is particularly true for local alignments lacking gaps, which we will consider first. Such alignments were precisely those sought by the original BLAST database search programs [6].

A local alignment without gaps consists simply of a pair of equal length segments, one from each of the two sequences being compared. A modification of the Smith-Waterman [7] or Sellers [8] algorithms will find all segment pairs whose scores can not be improved by extension or trimming. These are called high-scoring segment pairs or HSPs.

To analyze how high a score is likely to arise by chance, a model of random sequences is needed. For proteins, the simplest model chooses the amino acid residues in a sequence independently, with specific background probabilities for the various residues. Additionally, the expected score for aligning a random pair of amino acid is required to be negative. Were this not the case, long alignments would tend to have high score independently of whether the segments aligned were related, and the statistical theory would break down.

Just as the sum of a large number of independent identically distributed (i.i.d) random variables tends to a normal distribution, the maximum of a large number of i.i.d. random variables tends to an extreme value distribution [9]. (We will elide the many technical points required to make this statement rigorous.) In studying optimal local sequence alignments, we are essentially dealing with the latter case [10,11]. In the limit of sufficiently large sequence lengths m and n, the statistics of HSP scores are characterized by two parameters, K and *lambda*. Most simply, the expected number of HSPs with score at least S is given by the formula

$$E = Kmn\,e^{-\lambda S} \qquad (1)$$

We call this the E-value for the score S.

This formula makes eminently intuitive sense. Doubling the length of either sequence should double the number of HSPs attaining a given score. Also, for an HSP to attain the score $2x$ it must attain the score x twice in a row, so one expects E to decrease exponentially with score. The parameters K and *lambda* can be thought of simply as natural scales for the search space size and the scoring system respectively.

Bit scores

Raw scores have little meaning without detailed knowledge of the scoring system used, or more simply its statistical parameters *K* and *lambda*. Unless the scoring system is understood, citing a raw score alone is like citing a distance without specifying feet, meters, or light years. By normalizing a raw score using the formula

$$S' = \frac{\lambda S - \ln K}{\ln 2} \qquad (2)$$

one attains a "bit score" *S'*, which has a standard set of units. The *E*-value corresponding to a given bit score is simply

$$E = mn\,2^{-S'} \qquad (3)$$

Bit scores subsume the statistical essence of the scoring system employed, so that to calculate significance one needs to know in addition only the size of the search space.

P-values

The number of random HSPs with score >= *S* is described by a Poisson distribution [10,11]. This means that the probability of finding exactly *a* HSPs with score >=*S* is given by

$$e^{-E}\frac{E^a}{a!} \qquad (4)$$

where *E* is the *E*-value of *S* given by equation (1) above. Specifically the chance of finding zero HSPs with score >=*S* is e^{-E}, so the probability of finding at least one such HSP is

$$P = 1 - e^{-E} \qquad (5)$$

This is the *P*-value associated with the score *S*. For example, if

one expects to find three HSPs with score >= S, the probability of finding at least one is 0.95. The BLAST programs report E-value rather than P-values because it is easier to understand the difference between, for example, E-value of 5 and 10 than P-values of 0.993 and 0.99995. However, when $E < 0.01$, P-values and E-value are nearly identical.

Database searches

The E-value of equation (1) applies to the comparison of two proteins of lengths m and n. How does one assess the significance of an alignment that arises from the comparison of a protein of length m to a database containing many different proteins, of varying lengths? One view is that all proteins in the database are *a priori* equally likely to be related to the query. This implies that a low E-value for an alignment involving a short database sequence should carry the same weight as a low E-value for an alignment involving a long database sequence. To calculate a "database search" E-value, one simply multiplies the pairwise-comparison E-value by the number of sequences in the database. Recent versions of the FASTA protein comparison programs [12] take this approach [13].

An alternative view is that a query is *a priori* more likely to be related to a long than to a short sequence, because long sequences are often composed of multiple distinct domains. If we assume the *a priori* chance of relatedness is proportional to sequence length, then the pairwise E-value involving a database sequence of length n should be multiplied by N/n, where N is the total length of the database in residues. Examining equation (1), this can be accomplished simply by treating the database as a single long sequence of length N. The BLAST programs [6,14,15] take this approach to calculating database E-value. Notice that for DNA sequence comparisons, the length of database records is largely arbitrary, and therefore this is the only really tenable method for estimating statistical significance.

The statistics of gapped alignments

The statistics developed above have a solid theoretical foundation only for local alignments that are not permitted to have gaps. However, many computational experiments [14-21] and some analytic results [22] strongly suggest that the same theory applies as well to gapped alignments. For ungapped alignments, the statistical parameters can be calculated, using analytic formulas, from the substitution scores and the background residue frequencies of the sequences being

compared. For gapped alignments, these parameters must be estimated from a large-scale comparison of "random" sequences.

Some database search programs, such as FASTA [12] or various implementation of the Smith-Waterman algorithm [7], produce optimal local alignment scores for the comparison of the query sequence to every sequence in the database. Most of these scores involve unrelated sequences, and therefore can be used to estimate *lambda* and *K* [17,21]. This approach avoids the artificiality of a random sequence model by employing real sequences, with their attendant internal structure and correlations, but it must face the problem of excluding from the estimation scores from pairs of related sequences. The BLAST programs achieve much of their speed by avoiding the calculation of optimal alignment scores for all but a handful of unrelated sequences. The must therefore rely upon a pre-estimation of the parameters *lambda* and *K*, for a selected set of substitution matrices and gap costs. This estimation could be done using real sequences, but has instead relied upon a random sequence model [14], which appears to yield fairly accurate results [21].

Edge effects

The statistics described above tend to be somewhat conservative for short sequences. The theory supporting these statistics is an asymptotic one, which assumes an optimal local alignment can begin with any aligned pair of residues. However, a high-scoring alignment must have some length, and therefore can not begin near to the end of either of two sequences being compared. This "edge effect" may be corrected for by calculating an "effective length" for sequences [14]; the BLAST programs implement such a correction. For sequences longer than about 200 residues the edge effect correction is usually negligible.

The choice of substitution scores

The results a local alignment program produces depend strongly upon the scores it uses. No single scoring scheme is best for all purposes, and an understanding of the basic theory of local alignment scores can improve the sensitivity of one's sequence analyses. As before, the theory is fully developed only for scores used to find ungapped local alignments, so we start with that case.

A large number of different amino acid substitution scores, based upon a variety of rationales, have been described [23-36].

However the scores of any substitution matrix with negative expected score can be written uniquely in the form

$$s_{ij} = \left(\ln \frac{q_{ij}}{p_i p_j}\right) / \lambda \qquad (6)$$

where the q_{ij}, called target frequencies, are positive numbers that sum to 1, the p_i are background frequencies for the various residues, and *lambda* is a positive constant [10,31]. The *lambda* here is identical to the *lambda* of equation (1).

Multiplying all the scores in a substitution matrix by a positive constant does not change their essence: an alignment that was optimal using the original scores remains optimal. Such multiplication alters the parameter *lambda* but not the target frequencies q_{ij}. Thus, up to a constant scaling factor, every substitution matrix is uniquely determined by its target frequencies. These frequencies have a special significance [10,31]:

> A given class of alignments is best distinguished from chance by the substitution matrix whose target frequencies characterize the class.

To elaborate, one may characterize a set of alignments representing homologous protein regions by the frequency with which each possible pair of residues is aligned. If valine in the first sequence and leucine in the second appear in 1% of all alignment positions, the target frequency for (valine, leucine) is 0.01. The most direct way to construct appropriate substitution matrices for local sequence comparison is to estimate target and background frequencies, and calculate the corresponding log-odds scores of formula (6). These frequencies in general can not be derived from first principles, and their estimation requires empirical input.

▸ The PAM and BLOSUM amino acid substitution matrices

While all substitution matrices are implicitly of log-odds form, the first explicit construction using formula (6) was by Dayhoff and coworkers [24,25]. From a study of observed residue

replacements in closely related proteins, they constructed the PAM (for "point accepted mutation") model of molecular evolution. One "PAM" corresponds to an average change in 1% of all amino acid positions. After 100 PAMs of evolution, not every residue will have changed: some will have mutated several times, perhaps returning to their original state, and others not at all. Thus it is possible to recognize as homologous proteins separated by much more than 100 PAMs. Note that there is no general correspondence between PAM distance and evolutionary time, as different protein families evolve at different rates.

Using the PAM model, the target frequencies and the corresponding substitution matrix may be calculated for any given evolutionary distance. When two sequences are compared, it is not generally known a priori what evolutionary distance will best characterize any similarity they may share. Closely related sequences, however, are relatively easy to find even will non-optimal matrices, so the tendency has been to use matrices tailored for fairly distant similarities. For many years, the most widely used matrix was PAM-250, because it was the only one originally published by Dayhoff.

Dayhoff's formalism for calculating target frequencies has been criticized [27], and there have been several efforts to update her numbers using the vast quantities of derived protein sequence data generated since her work [33,35]. These newer PAM matrices do not differ greatly from the original ones [37].

An alternative approach to estimating target frequencies, and the corresponding log-odds matrices, has been advanced by Henikoff & Henikoff [34]. They examine multiple alignments of distantly related protein regions directly, rather than extrapolate from closely related sequences. An advantage of this approach is that it cleaves closer to observation; a disadvantage is that it yields no evolutionary model. A number of tests [13,37] suggest that the "BLOSUM" matrices produced by this method generally are superior to the PAM matrices for detecting biological relationships.

▸ DNA substitution matrices

While we have discussed substitution matrices only in the context of protein sequence comparison, all the main issues carry over to DNA sequence comparison. One warning is that when the sequences of interest code for protein, it is almost always better to compare the protein translations than to compare the DNA sequences directly. The reason is that after only a small amount of evolutionary change, the DNA

sequences, when compared using simple nucleotide substitution scores, contain less information with which to deduce homology than do the encoded protein sequences [32].

Sometimes, however, one may wish to compare non-coding DNA sequences, at which point the same log-odds approach as before applies. An evolutionary model in which all nucleotides are equally common and all substitution mutations are equally likely yields different scores only for matches and mismatches [32]. A more complex model, in which transitions are more likely than transversions, yields different "mismatch" scores for transitions and transversions [32]. The best scores to use will depend upon whether one is seeking relatively diverged or closely related sequences [32].

Gap scores

Our theoretical development concerning the optimality of matrices constructed using equation (6) unfortunately is invalid as soon as gaps and associated gap scores are introduced, and no more general theory is available to take its place. However, if the gap scores employed are sufficiently large, one can expect that the optimal substitution scores for a given application will not change substantially. In practice, the same substitution scores have been applied fruitfully to local alignments both with and without gaps. Appropriate gap scores have been selected over the years by trial and error [13], and most alignment programs will have a default set of gap scores to go with a default set of substitution scores. If the user wishes to employ a different set of substitution scores, there is no guarantee that the same gap scores will remain appropriate. No clear theoretical guidance can be given, but "affine gap scores" [38-41], with a large penalty for opening a gap and a much smaller one for extending it, have generally proved among the most effective.

Low complexity sequence regions

There is one frequent case where the random models and therefore the statistics discussed here break down. As many as one fourth of all residues in protein sequences occur within regions with highly biased amino acid composition. Alignments of two regions with similarly biased composition may achieve very high scores that owe virtually nothing to residue order but are due instead to segment composition. Alignments of such "low complexity" regions have little meaning in any case: since these regions most likely arise by gene slippage, the one-to-one residue correspondence imposed by alignment is not valid.

While it is worth noting that two proteins contain similar low complexity regions, they are best excluded when constructing alignments [42-44]. The BLAST programs employ the SEG algorithm [43] to filter low complexity regions from proteins before executing a database search.

References

[1] Fitch, W.M. (1983) "Random sequences." J. Mol. Biol. 163:171-176. (PubMed)

[2] Lipman, D.J., Wilbur, W.J., Smith T.F. & Waterman, M.S. (1984) "On the statistical significance of nucleic acid similarities." Nucl. Acids Res. 12:215-226. (PubMed)

[3] Altschul, S.F. & Erickson, B.W. (1985) "Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage." Mol. Biol. Evol. 2:526-538. (PubMed)

[4] Deken, J. (1983) "Probabilistic behavior of longest-common-subsequence length." In "Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison." D. Sankoff & J.B. Kruskal (eds.), pp. 55-91, Addison-Wesley, Reading, MA.

[5] Reich, J.G., Drabsch, H. & Daumler, A. (1984) "On the statistical assessment of similarities in DNA sequences." Nucl. Acids Res. 12:5529-5543. (PubMed)

[6] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410. (PubMed)

[7] Smith, T.F. & Waterman, M.S. (1981) "Identification of common molecular subsequences." J. Mol. Biol. 147:195-197. (PubMed)

[8] Sellers, P.H. (1984) "Pattern recognition in genetic sequences by mismatch density." Bull. Math. Biol. 46:501-514.

[9] Gumbel, E. J. (1958) "Statistics of extremes." Columbia University Press, New York, NY.

[10] Karlin, S. & Altschul, S.F. (1990) "Methods for assessing the statistical significance of molecular sequence features by using

general scoring schemes." Proc. Natl. Acad. Sci. USA 87:2264-2268.(PubMed)

[11] Dembo, A., Karlin, S. & Zeitouni, O. (1994) "Limit distribution of maximal non-aligned two-sequence segmental score." Ann. Prob. 22:2022-2039.

[12] Pearson, W.R. & Lipman, D.J. (1988) Improved tools for biological sequence comparison." Proc. Natl. Acad. Sci. USA 85:2444-2448. (PubMed)

[13] Pearson, W.R. (1995) "Comparison of methods for searching protein sequence databases." Prot. Sci. 4:1145-1160. (PubMed)

[14] Altschul, S.F. & Gish, W. (1996) "Local alignment statistics." Meth. Enzymol. 266:460-480. (PubMed)

[15] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res. 25:3389-3402. (PubMed)

[16] Smith, T.F., Waterman, M.S. & Burks, C. (1985) "The statistical distribution of nucleic acid similarities." Nucleic Acids Res. 13:645-656. (PubMed)

[17] Collins, J.F., Coulson, A.F.W. & Lyall, A. (1988) "The significance of protein sequence similarities." Comput. Appl. Biosci. 4:67-71. (PubMed)

[18] Mott, R. (1992) "Maximum-likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores." Bull. Math. Biol. 54:59-75.

[19] Waterman, M.S. & Vingron, M. (1994) "Rapid and accurate estimates of statistical significance for sequence database searches." Proc. Natl. Acad. Sci. USA 91:4625-4628. (PubMed)

[20] Waterman, M.S. & Vingron, M. (1994) "Sequence comparison significance and Poisson approximation." Stat. Sci. 9:367-381.

[21] Pearson, W.R. (1998) "Empirical statistical estimates for sequence similarity searches." J. Mol. Biol. 276:71-84. (PubMed)

[22] Arratia, R. & Waterman, M.S. (1994) "A phase transition for the score in matching random sequences allowing deletions." Ann. Appl. Prob. 4:200-225.

[23] McLachlan, A.D. (1971) "Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c-551." J. Mol. Biol. 61:409-424. (PubMed)

[24] Dayhoff, M.O., Schwartz, R.M. & Orcutt, B.C. (1978) "A model of evolutionary change in proteins." In "Atlas of Protein Sequence and Structure," Vol. 5, Suppl. 3 (ed. M.O. Dayhoff), pp. 345-352. Natl. Biomed. Res. Found., Washington, DC.

[25] Schwartz, R.M. & Dayhoff, M.O. (1978) "Matrices for detecting distant relationships." In "Atlas of Protein Sequence and Structure," Vol. 5, Suppl. 3 (ed. M.O. Dayhoff), p. 353-358. Natl. Biomed. Res. Found., Washington, DC.

[26] Feng, D.F., Johnson, M.S. & Doolittle, R.F. (1984) "Aligning amino acid sequences: comparison of commonly used methods." J. Mol. Evol. 21:112-125. (PubMed)

[27] Wilbur, W.J. (1985) "On the PAM matrix model of protein evolution." Mol. Biol. Evol. 2:434-447. (PubMed)

[28] Taylor, W.R. (1986) "The classification of amino acid conservation." J. Theor. Biol. 119:205-218. (PubMed)

[29] Rao, J.K.M. (1987) "New scoring matrix for amino acid residue exchanges based on residue characteristic physical parameters." Int. J. Peptide Protein Res. 29:276-281.

[30] Risler, J.L., Delorme, M.O., Delacroix, H. & Henaut, A. (1988) "Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix." J. Mol. Biol. 204:1019-1029. (PubMed)

[31] Altschul, S.F. (1991) "Amino acid substitution matrices from an information theoretic perspective." J. Mol. Biol. 219:555-565. (PubMed)

[32] States, D.J., Gish, W. & Altschul, S.F. (1991) "Improved sensitivity of nucleic acid database searches using application-specific scoring matrices." Methods 3:66-70.

[33] Gonnet, G.H., Cohen, M.A. & Benner, S.A. (1992)

"Exhaustive matching of the entire protein sequence database." Science 256:1443-1445. (PubMed)

[34] Henikoff, S. & Henikoff, J.G. (1992) "Amino acid substitution matrices from protein blocks." Proc. Natl. Acad. Sci. USA 89:10915-10919. (PubMed)

[35] Jones, D.T., Taylor, W.R. & Thornton, J.M. (1992) "The rapid generation of mutation data matrices from protein sequences." Comput. Appl. Biosci. 8:275-282. (PubMed)

[36] Overington, J., Donnelly, D., Johnson M.S., Sali, A. & Blundell, T.L. (1992) "Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds." Prot. Sci. 1:216-226. (PubMed)

[37] Henikoff, S. & Henikoff, J.G. (1993) "Performance evaluation of amino acid substitution matrices." Proteins 17:49-61. (PubMed)

[38] Gotoh, O. (1982) "An improved algorithm for matching biological sequences." J. Mol. Biol. 162:705-708. (PubMed)

[39] Fitch, W.M. & Smith, T.F. (1983) "Optimal sequence alignments." Proc. Natl. Acad. Sci. USA 80:1382-1386.

[40] Altschul, S.F. & Erickson, B.W. (1986) "Optimal sequence alignment using affine gap costs." Bull. Math. Biol. 48:603-616. (PubMed)

[41] Myers, E.W. & Miller, W. (1988) "Optimal alignments in linear space." Comput. Appl. Biosci. 4:11-17. (PubMed)

[42] Claverie, J.-M. & States, D.J. (1993) "Information enhancement methods for large-scale sequence-analysis." Comput. Chem. 17:191-201.

[43] Wootton, J.C. & Federhen, S. (1993) "Statistics of local complexity in amino acid sequences and sequence databases." Comput. Chem. 17:149-163.

[44] Altschul, S.F., Boguski, M.S., Gish, W. & Wootton, J.C. (1994) "Issues in searching molecular sequence databases." Nature Genet. 6:119-129. (PubMed)

Comparison in similarity of 16S rDNA of Corynebacterium glutamicum with microorganisms belonging to the genus Microbacterium or Brevibacterium

Similarities of the nucleotide sequences of 16S rDNA registered in GenBank shown below were compared.

Corynebacterium glutamicum ATCC 13032□GenBank accession No. AP05274
Microbacterium lacticum□GenBank accession No. AB007415
Brevibacterium linens DSM 20425□GenBank accession No. X77451

Table: 16S rDNA Similarity (% similarity)

	C.glutamicum	M.lacticum	B.linens
Corynebacterium glutamicum	100.0	86.8	87.8
Microbacterium lacticum		100.0	87.3
Brevibacterium linens			100.0

Files in GenBank of the sequences used are shown below.

LOCUS AP005274 1524 bp DNA linear BCT 08-AUG-2002
DEFINITION Corynebacterium glutamicum ATCC 13032 DNA, complete genome, section
1/10.
ACCESSION AP005274 REGION: 76643..78166
VERSION AP005274.1 GI:21322764
KEYWORDS .
SOURCE Corynebacterium glutamicum ATCC 13032
ORGANISM Corynebacterium glutamicum ATCC 13032
Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales;
Corynebacterineae; Corynebacteriaceae; Corynebacterium.
REFERENCE 1
AUTHORS Nakagawa,S.
TITLE Complete genomic sequence of Corynebacterium glutamicum ATCC 13032
JOURNAL Unpublished
REFERENCE 2 (bases 1 to 1524)
AUTHORS Nakagawa,S.
TITLE Direct Submission
JOURNAL Submitted (24-MAY-2002) Satoshi Nakagawa, Kyowa Hakko Kogyo Co.
Ltd., Tokyo Research Laboratories; 3-6-6, Asahi-machi, Machida,
Tokyo 194-8533, Japan (E-mail:snakagawa@xanagen.com,
Tel:81-44-829-3031, Fax:81-44-813-1651)
COMMENT This sequence is conducted by collaboration of Kyowa Hakko Kogyo
Co. Ltd. And Kitasato University.
FEATURES Location/Qualifiers
source 1..1524
/organism="Corynebacterium glutamicum ATCC 13032"
/mol_type="genomic DNA"
/strain="ATCC 13032"
/db_xref="taxon:196627"
/note="ATCC 13032"
rRNA 1..1524

/product="16S ribosomal RNA"
BASE COUNT 339 a 355 c 506 g 324 t
ORIGIN
1 tgtggagagt ttgatcctgg ctcaggacga acgctggcgg cgtgcttaac acatgcaagt
61 cgaacgctga aaccagagct tgctttggtg gatgagtggc gaacgggtga gtaacacgtg
121 ggtgatctgc cctacacttt gggataagcc tgggaaactg ggtctaatac cgaatattca
181 caccaccgta ggggtggtgt ggaaagcttt atgcggtgtg ggatgagcct gcggcctatc
241 agcttgttgg tggggtaatg gcctaccaag gcgtcgacgg gtagccggcc tgagagggtg
301 tacggccaca ttgggactga gacacggccc agactcctac gggaggcagc agtggggaat
361 attgcacaat gggcgcaagc ctgatgcagc gacgccgcgt gggggatgaa ggccttcggg
421 ttgtaaactc ctttcgctag ggacgaagcc ttatggtgac ggtacctgga gaagaagcac
481 cggctaacta cgtgccagca gccgcggtaa tacgtagggt gcgagcgttg tccggaatta
541 ctgggcgtaa agagctcgta ggtggtttgt cgcgtcgtct gtgaaatccc ggggcttaac
601 ttcgggcgtg caggcgatac gggcataact tgagtgctgt aggggagact ggaattcctg
661 gtgtagcggt gaaatgcgca gatatcagga ggaacaccaa tggcgaaggc aggtctctgg
721 gcagtaactg acgctgagga gcgaaagcat gggtagcgaa caggattaga taccctggta
781 gtccatgccg taaacggtgg gcgctaggtg taggggtctt ccacgacttc tgtgccgcag
841 ctaacgcatt aagcgccccg cctggggagt acggccgcaa ggctaaaact caaaggaatt
901 gacgggggcc cgcacaagcg gcggagcatg tggattaatt cgatgcaacg cgaagaacct
961 tacctgggct tgacatggac cggatcggcg tagagatacg ttttcccttg tggtcggttc
1021 acaggtggtg catggttgtc gtcagctcgt gtcgtgagat gttgggttaa gtcccgcaac
1081 gagcgcaacc cttgtcttat gttgccagca cattgtggtg ggtactcatg agagactgcc
1141 ggggttaact cggaggaagg tggggatgac gtcaaatcat catgcccctt atgtccaggg
1201 cttcacacat gctacaatgg tcggtacagc gagttgccac accgtgaggt ggagctaatc
1261 tcttaaagcc ggcctcagtt cggattgggg tctgcaactc gaccccatga agtcggagtc
1321 gctagtaatc gcagatcagc aacgctgcgg tgaatacgtt cccgggcctt gtacacaccg
1381 cccgtcacgt catgaaagtt ggtaacaccc gaagccagtg gcccaacctt ttaggggga
1441 gctgtcgaag gtgggatcgg cgattgggac gaagtcgtaa caaggtagcc gtaccggaag
1501 gtgcggctgg atcacctcct ttct
//

LOCUS AB007415 1456 bp rRNA linear BCT 27-APR-1999
DEFINITION Microbacterium lacticum 16S rRNA, partial sequence.
ACCESSION AB007415
VERSION AB007415.1 GI:4648292
KEYWORDS 16S ribosomal RNA.
SOURCE Microbacterium lacticum
ORGANISM Microbacterium lacticum
Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales;
Micrococcineae; Microbacteriaceae; Microbacterium.
REFERENCE 1 (sites)
AUTHORS Takeuchi,M. and Hatano,K.
TITLE Proposal of six new species in the genus Microbacterium and
transfer of Flavobacterium marinotypicum ZoBell and Upham to the
genus Microbacterium as Microbacterium maritypicum comb. nov
JOURNAL Int. J. Syst. Bacteriol. 48 Pt 3, 973-982 (1998)
MEDLINE 98404565
PUBMED 9734054
REFERENCE 2 (bases 1 to 1456)
AUTHORS Takeuchi,M.
TITLE Direct Submission
JOURNAL Submitted (11-SEP-1997) Mariko Takeuchi, Institute for
Fermentation, Osaka, Bacteria G; 2-17-85, Juso-honmachi,
Yodogawa-ku, Yodogawa-ku, Osaka 532, Japan
(E-mail:fvgg0814@mb.infoweb.ne.jp, Tel:06-6300-6310,
Fax:06-6300-6814)
FEATURES Location/Qualifiers

```
   source          1..1456
                   /organism="Microbacterium lacticum"
                   /mol_type="rRNA"
                   /db_xref="taxon:33885"
   rRNA            <1..>1456
                   /product="16S ribosomal RNA"
BASE COUNT     356 a    350 c    468 g    282 t
ORIGIN
      1 agtttgatcc tggctcagga tgaacgctgg cggcgtgctt aacacatgca agtcgaacgg
     61 tgaagcggag ctatgctctg ctggatagtg gcggaacggg tgagtaacac gtgagcaatc
    121 tgcccctgac tctgggataa gcgctggaaa cggcgtctaa tactggatac gagctgcgaa
    181 ggcatcttca gcagctggaa agaacttcgg tcagggatga gctcgcggcc tatcagcttg
    241 ttggtgaggt aatggctcac caaggcgtcg acgggtagcc ggcctgagag ggtgaccggc
    301 cacactggga ctgagacacg gcccagactc ctacgggagg cagcagtggg gaatattgca
    361 caatgggcga aagcctgatg cagcaacgcc gcgtgaggga cgacggcctt cgggttgtaa
    421 acctctttta gcagggaaga agcgaaagtg acggtacctg cagaaaaagc gccggctaac
    481 tacgtgccag cagccgcggt aatacgtagg gcgcaagcgt tatccggaat tattgggcgt
    541 aaagagctcg taggcggttt gtcgcgtctg ctgtgaaatc ccgaggctca acctcgggcc
    601 tgcagtgggt acgggcagac tagagtgcgg taggggagat tggaattcct ggtgtagcgg
    661 tggaatgcgc agatatcagg aggaacaccg atggcgaagg cagatctctg ggccgtaact
    721 gacgctgagg agcgaaaggg tggggagcaa acaggcttag ataccctggt agtccacccc
    781 gtaaacgttg ggaactagtt gtggggtcca ttccacggat tccgtgacgc agctaacgca
    841 ttaagttccc cgcctgggga gtacggccgc aaggctaaaa ctcaaaggaa ttgacgggga
    901 cccgcacaag cggcggagca tgcggattaa ttcgatgcaa cgcgaagaac cttaccaagg
    961 cttgacatat acgagaacgg gccagaaatg gtcaactctt tggacactcg taaacaggtg
   1021 gtgcatggtt gtcgtcagct cgtgtcgtga gatgttgggt taagtcccgc aacgagcgca
   1081 accctcgttc tatgttgcca gcacgtaatg gtgggaactc atggaatact gccggggtca
   1141 actcggagga aggtgaggat gacgtcaaat catcatgccc cttatgtctt gggcttcacg
   1201 catgctacaa tggccggtac aatgggctgc aataccgcaa ggtggagcga atcccaaaaa
   1261 gccggtccca gttcggattg aggtctgcaa ctcgacctca tgaagtcgga gtcgctagta
   1321 atcgcagatc agcaacgctg cggtgaatac gttcccgggt cttgtacaca ccgcccgtca
   1381 agtcatgaaa gtcggtaaca cctgaagccg gtggcccaac ccttgtggag ggagccgtcg
   1441 aaggtgggat ccgtaa
//


LOCUS       BL16SR              1473 bp    DNA     linear   BCT 19-JUL-1994
DEFINITION  B.linens (DSM 20425) 16S rRNA gene.
ACCESSION   X77451
VERSION     X77451.1  GI:515022
KEYWORDS    16S ribosomal RNA.
SOURCE      Brevibacterium linens
  ORGANISM  Brevibacterium linens
            Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales;
            Micrococcineae; Brevibacteriaceae; Brevibacterium.
REFERENCE   1
  AUTHORS   Rainey,F.A., Weiss,N., Prauser,H. and Stackebrandt,E.
  TITLE     Further evidence for the phylogenetic coherence of actinomycetes
            with group B-peptidoglycan
  JOURNAL   FEMS Microbiol. Lett. 118, 135-140 (1994)
REFERENCE   2  (bases 1 to 1473)
  AUTHORS   Stackebrandt,E.
  TITLE     Direct Submission
  JOURNAL   Submitted (31-JAN-1994) E. Stackebrandt, DSM Deutsche Sammlung
von
            Mikroorganis. und Zellkulturen GmbH, Mascheroder Weg 1B, 38124
            Braunschweig, FRG
FEATURES             Location/Qualifiers
```

```
     source          1..1473
                     /organism="Brevibacterium linens"
                     /mol_type="genomic DNA"
                     /strain="DSM 20425"
                     /db_xref="taxon:1703"
     rRNA            1..1473
                     /product="16S ribosomal RNA"
BASE COUNT     341 a    350 c    485 g    295 t      2 others
ORIGIN
        1 gacgaacgct ggctgcgtgc ttaacacatg caagtcgaac gctgaaccag gagcttgctc
       61 tgtggatgag tggcgaacgg gtgagtaaca cgtgagtaac ctgcccccga tttcgggata
      121 agcccgggaa actgggtcta ataccggata cgaccaatcc tcgcatgagg gttggtggaa
      181 agtttttcga tcggggatgg gctcgcggcc tatcagcttg ttggtggggt aatggcctac
      241 caagccgacg acgggtagcc ggcctgagag ggcgaccggc cacactggga ctgagacacg
      301 gcccagactc ctacgggagg cagcagtggg gaatattgca caatggggga aaccctgatg
      361 cagcgacgca gcgtgcggga tgacggcctt cgggttgtaa accgctttca gcagggaaga
      421 agccgtaagt gacggtacct ccagaagaag taccggctaa ctacgtgcca gcagccgcgg
      481 taatacgtag ggtacgagcg ttgtccggaa ttattgggcg taaagagctc gtaggtggnt
      541 gctcacgtct gctgtggaaa cgcgacgctt aacgttgcgc gtgcagtggg tacgggctga
      601 ctagagtgca gtaggggagt ctggaattcc tggtgtagcg gtgaaatgcg cagatatcag
      661 gaggaacacc ggtgncgaag gcgggactct gggctgtaac tgacactgag gagcgaaagc
      721 atggggagcg aacaggatta gataccctgg tagtccatgc cgtaaacgtt gggcactagg
      781 tgtgggggac attccacgtt ctccgcgccg tagctaacgc attaagtgcc ccgcctgggg
      841 agtacggtcg caaggctaaa actcaaagga attgacgggg gcccgcacaa gcggcggagc
      901 atgcggatta attcgatgca acgcgaagaa ccttaccaag gcttgacata cactggaccg
      961 ttctggaaac agttctctct ttggagctgg tgtacaggtg gtgcatggtt gtcgtcagct
     1021 cgtgtcgtga gatgttgggt taagtcccgc aacgagcgca accctcgttc tatgttgcca
     1081 gcacgtgatg gtgggaactc ataggagact gccggggtca actcggagga aggtggggat
     1141 gacgtcaaat catcatgccc tttatgtctt gggcttcacg catgctacaa tggctggtac
     1201 agagagaggc gaacccgtga gggtgagcga atcccttaaa gccagtctca gttcggatcg
     1261 tagtctgcaa ttcgactacg tgaagtcgga gtcgctagta atcgcagatc agcaacgctg
     1321 cggtgaatac gttcccgggc cttgtacaca ccgcccgtca agtcacgaaa gtcggtaaca
     1381 cccgaagccg gtgtcccaac ccccttgtg ggaggggcg tctaggtggg actggtgatt
     1441 gggactaagt cgtaacaagg tagccgtacc gga
//
```

(32) The method according to any one of (24), (26), (28) and (30), wherein a coryneform bacterium is a microorganism of the genus *Corynebacterium*, the genus *Brevibacterium*, or the genus *Microbacterium*.

(33) The system according to (31), wherein the microorganism belonging to the genus *Corynebacterium* is selected from the group consisting of *Corynebacterium glutamicum*, *Corynebacterium acetoacidophilum*, *Corynebacterium acetoglutamicum*, *Corynebacterium callunae*, *Corynebacterium herculis*, *Corynebacterium lilium*, *Corynebacterium melassecola*, *Corynebacterium thermoaminogenes*, and *Corynebacterium ammoniagenes*.

(34) The method according to (32), wherein the microorganism belonging to the genus *Corynebacterium* is selected from the group consisting of *Corynebacterium glutamicum*, *Corynebacterium acetoacidophilum*, *Corynebacterium acetoglutamicum*, *Corynebacterium callunae*, *Corynebacterium herculis*, *Corynebacterium lilium*, *Corynebacterium melassecola*, *Corynebacterium thermoaminogenes*, and *Corynebacterium ammoniagenes*.

(35) A recording medium or storage device which is readable by a computer in which at least one nucleotide sequence information selected from SEQ ID NOS:1 to 3501 or function information based on the nucleotide sequence is recorded, and is usable in the system of (23) or (27) or the method of (24) or (28).

(36) A recording medium or storage device which is readable by a computer in which at least one amino acid sequence information selected from SEQ ID NOS:3502 to 7001 or function information based on the amino acid sequence is recorded, and is usable in the system of (25) or (29) or the method of (26) or (30).

(37) The recording medium or storage device according to (35) or (36), which is a computer readable recording medium selected from the group consisting of a floppy disc, a hard disc, a magnetic tape, a random access memory (RAM), a read only memory (ROM), a magneto-optic disc (MO), CD-ROM, CD-R, CD-RW, DVD-ROM, DVD-RAM and DVD-RW.

(38) A polypeptide having a homoserine dehydrogenase activity, comprising an amino acid sequence in which the Val residue at the 59th in the amino acid sequence of homoserine dehydrogenase derived from a coryneform bacterium is replaced with an amino acid residue other than a Val residue.

(39) A polypeptide comprising an amino acid sequence in which the Val residue at the 59th position in the amino acid sequence as represented by SEQ ID NO:6952 is replaced with an amino acid residue other than a Val residue.

(40) The polypeptide according to (38) or (39), wherein the Val residue at the 59th position is replaced with an Ala residue.

(41) A polypeptide having pyruvate carboxylase activity, comprising an amino acid sequence in which the Pro residue at the 458th position in the amino acid sequence of pyruvate carboxylase derived from a coryneform bacterium is replaced with an amino acid residue other than a Pro residue.

(42) A polypeptide comprising an amino acid sequence in which the Pro residue at the 458th position in the amino acid sequence represented by SEQ ID NO:4265 is replaced with an amino acid residue other than a Pro residue.

(43) The polypeptide according to (41) or (42), wherein the Pro residue at the 458th position is replaced with a Ser residue.

(44) The polypeptide according to any one of (38) to (43), which is derived from *Corynebacterium glutamicum*.

(45) A DNA encoding the polypeptide of any one of (38) to (44).

(46) A recombinant DNA comprising the DNA of (45).

(47) A transformant comprising the recombinant DNA of (46).

(48) A transformant comprising in its chromosome the DNA of (45).

(49) The transformant according to (47) or (48), which is derived from a coryneform bacterium.

(50) The transformant according to (49), which is derived from *Corynebacterium glutamicum*.

(51) A method for producing L-lysine, comprising:

culturing the transformant of any one of (47) to (50) in a medium to produce and accumulate L-lysine in the medium, and

recovering the L-lysine from the culture.

(52) A method for breeding a coryneform bacterium using the nucleotide sequence information represented by SEQ ID NOS:1 to 3431, comprising the following:

(i) comparing a nucleotide sequence of a genome or gene of a production strain derived a coryneform bacterium which has been subjected to mutation breeding so as to produce at least one compound selected from an amino acid, a nucleic acid, a vitamin, a saccharide, an organic acid, and analogous thereof by a fermentation method, with a corresponding nucleotide sequence in SEQ ID NOS:1 to 3431;

(ii) identifying a mutation point present in the production strain based on a result obtained by (i);

(iii) introducing the mutation point into a coryneform bacterium which is free of the mutation point; and

(iv) examining productivity by the fermentation method of the compound selected in (i) of the coryneform bacterium obtained in (iii).

(53) The method according to (52), wherein the gene is a gene encoding an enzyme in a biosynthetic pathway or a signal transmission pathway.

(54) The method according to (52), wherein the mutation point is a mutation point relating to a useful mutation which improves or stabilizes the productivity.

(55) A method for breading a coryneform bacterium using the nucleotide sequence information represented by SEQ ID NOS:1 to 3431, comprising:

(i) comparing a nucleotide sequence of a genome or gene of a production strain derived a coryneform bacterium which has been subjected to mutation breeding so as to produce at least one compound selected from an amino acid, a nucleic acid, a vitamin, a saccharide, an organic acid, and analogous thereof by a fermentation method, with a corresponding nucleotide sequence in SEQ ID NOS:1 to 3431;

(ii) identifying a mutation point present in the production strain based on a result obtain by (i);

(iii) deleting a mutation point from a coryneform bacterium having the mutation point; and

(iv) examining productivity by the fermentation method of the compound selected in (i) of the coryneform bacterium obtained in (iii).

(56) The method according to (55), wherein the gene is a gene encoding an enzyme in a biosynthetic pathway or a signal transmission pathway.

(57) The method according to (55), wherein the mutation point is a mutation point which decreases or destabilizes the productivity.

(58) A method for breeding a coryneform bacterium using the nucleotide sequence information represented by SEQ ID NOS:2 to 3431, comprising the following:

(i) identifying an isozyme relating to biosynthesis of at least one compound selected from an amino acid, a nucleic acid, a vitamin, a saccharide, an organic acid, and analogous thereof, based on the nucleotide sequence information represented by SEQ ID NOS:2 to 3431;

(ii) classifying the isozyme identified in (i) into an isozyme having the same activity;

(iii) mutating all genes encoding the isozyme having the same activity simultaneously; and

(iv) examining productivity by a fermentation method of the compound selected in (i) of the coryneform bacterium which have been transformed with the gene obtained in (iii).

(59) A method for breeding a coryneform bacterium using the nucleotide sequence information represented by SEQ ID NOS:2 to 3431, comprising the following:

(i) arranging a function information of an open reading frame (ORF) represented by SEQ ID NOS:2 to 3431;

(ii) allowing the arranged ORF to correspond to an enzyme on a known biosynthesis or signal transmission pathway;

(iii) explicating an unknown biosynthesis pathway or signal transmission pathway of a coryneform bacterium in combination with information relating known biosynthesis

pathway or signal transmission pathway of a coryneform bacterium;

(iv) comparing the pathway explicated in (iii) with a biosynthesis pathway of a target useful product; and

(v) transgenetically varying a coryneform bacterium based on the nucleotide sequence information to either strengthen a pathway which is judged to be important in the biosynthesis of the target useful product in (iv) or weaken a pathway which is judged not to be important in the biosynthesis of the target useful product in (iv).

(60) A coryneform bacterium, bred by the method of any one of (52) to (59).

(61) The coryneform bacterium according to (60), which is a microorganism belonging to the genus *Corynebacterium*, the genus *Brevibacterium*, or the genus *Microbacterium*.

(62) The coryneform bacterium according to (61), wherein the microorganism belonging to the genus *Corynebacterium* is selected from the group consisting of *Corynebacterium glutamicum*, *Corynebacterium acetoacidophilum*, *Corynebacterium acetoglutamicum*, *Corynebacterium callunae*, *Corynebacterium herculis*, *Corynebacterium lilium*, *Corynebacterium melassecola*, *Corynebacterium thermoaminogenes*, and *Corynebacterium ammoniagenes*.

(63) A method for producing at least one compound selected from an amino acid, a nucleic acid, a vitamin, a saccharide, an organic acid and an analogue thereof, comprising: